Knowledge-Based Systems 24 (2011) 386–392

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Interestingness measures for association rules based on statistical validity

Izwan Nizal Mohd. Shaharanee *, Fedja Hadzic, Tharam S. Dillon

*Digital Ecosystem and Business Intelligence Institute, Curtin University of Technology, Perth 6102, Australia*

ARTICLE INFO

*Article history:*
Received 14 June 2010
Received in revised form 6 October 2010
Accepted 22 November 2010
Available online 28 November 2010

*Keywords:*
Data mining
Structured data
Interesting rules
Statistical analysis
Redundant rules
Interestingness measure

ABSTRACT

Assessing rules with interestingness measures is the pillar of successful application of association rules discovery. However, association rules discovered are normally large in number, some of which are not considered as interesting or significant for the application at hand. In this paper, we present a systematic approach to ascertain the discovered rules, and provide a precise statistical approach supporting this framework. The proposed strategy combines data mining and statistical measurement techniques, including redundancy analysis, sampling and multivariate statistical analysis, to discard the non- significant rules. Moreover, we consider real world datasets which are characterized by the uniform and non-uniform data/items distribution with a mixture of measurement levels throughout the data/items. The proposed unified framework is applied on these datasets to demonstrate its effectiveness in discarding many of the redundant or non-significant rules, while still preserving the high accuracy of the rule set as a whole.



## 1. Introduction

Data mining or knowledge discovery from data (KDD) is known for its capabilities in offering systematic ways of acquiring useful rules and patterns from large quantities of data. The rules derived from data mining application are considered interesting and useful if they are comprehensible, valid on tests and new unseen data with an appropriate degree of certainty, potentially useful, actionable, and novel [17]. McGarry [24] claims that the majority of data mining/machine learning type patterns are rule based in nature with a well defined structure, such as rules derived from decision trees and association rules. The most common patterns that can be evaluated by interestingness measures include association rules, classification rules, and summaries [14]. Association rule mining is one of the most popular data mining techniques widely used for discovering interesting associations and correlations between data elements in a diverse range of applications [34]. The association rule mining techniques may differ from one another, but a common feature is that all the frequent patterns are first extracted and then association rule are formed from such patterns. Frequent pattern extraction plays an important part in generating good and interesting rules, and is considered as the most difficult and complex task.

Our work in the area of rules interestingness measures is motivated by the objective interestingness measures which are based on probability theory, statistics and information theory. Various objective interestingness criteria have been used to limit the nature of rules extracted, as explained in [14]. A number of researchers have anticipated an assessment of pattern discovery by applying a statistical significance test as discussed in [16,19,20,29,30,32].

Assessing whether a rule satisfies a particular constraint is accompanied by a risk that the rule will satisfy the constraint with respect to the sample data but not with respect to the whole data distribution [30]. As such, the rules may not reflect the "real" association between the underlying attributes. The hypotheses reflected in the generated rules must be validated by a statistical methodology in order for them to be useful in practice, because the nature of data mining techniques is data driven [15]. However, even if the rules satisfy appropriate statistical tests, the underlying association may nevertheless be caused purely by a statistical coincidence [5].

The contribution of the work presented in this paper, is the development of systematic ways to verify the usefulness of rules obtained from association rules mining using statistical analysis. A unified framework is proposed that combines several techniques to access the quality of rules, and removes any redundant and unnecessary rules. Initial ideas and preliminary results were presented earlier in [26,27]. Several extensions and refinements were made so that the method could be applied to more realistic datasets including complex data types, infrequent items and uneven attribute value distribution. Furthermore, a comparison of the statistical measure used in our framework with the popular Mutual Information measure is included. The rest of the paper is organized as follows. Section 2, briefly overviews the problem of ascertaining

* Corresponding author. Tel.: +61 892669270; fax: +61 8 9266 7548.
*E-mail addresses:* izwan.mohdshaharanee@postgrad.curtin.edu.au (Izwan Nizal Mohd. Shaharanee), f.hadzic@cbs.curtin.edu.au (F. Hadzic), tharam.dillon@cbs.curtin.edu.au (T.S. Dillon).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.11.005

the discovered rules, followed by related work in Section 3. In Section 4, we describe our proposed framework and its formal definition. The framework is evaluated using real world datasets and several experimental findings and an explanation are given in Section 5. Section 6 concludes the paper and describes our ongoing work in this field of study.

## 2. Problem definitions

Association rule mining in its most fundamental structure is used to discover interesting relationships among items in a given dataset under minimum support and confidence conditions. A commonly used example is in market basket analysis, where an association rule $X \rightarrow Y$ means if a consumer buys the set of items $X$, then he/she probably also buys items $Y$. These items are typically referred to as itemsets. The problem of finding association rules $X \rightarrow Y$ was first introduced in [1,2] as a data mining task of finding frequently co-occurring items in a large Boolean transaction database. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Each transaction $T$ is a set of items, such that $T \subseteq I$. An association rule is a condition of the form of $X \rightarrow Y$ where $X \subseteq I$ and $Y \subseteq I$ are two sets of items. The support of a rule $X \rightarrow Y$ is the number of transactions that contain both $X$ and $Y$, while the confidence of a rule $X \rightarrow Y$ is the number of transactions containing $X$, that also contain $Y$.

Bing et al. [9], Freitas [13] and Yun et al. [33] argue that for a real large database that is often comprised of either relatively frequent/infrequent items, using multiple and relative support should be considered. The rules satisfying the standard support and confidence constraints are often too numerous to be utilized efficiently and effectively for the application at hand [21]. Many patterns from the frequent pattern set are often redundant and unrelated [31]. Webb [30] defines redundant rules as those rules that include items in the antecedent that are entailed by the other elements of the antecedents. Redundant rule constraints discard rule $x \rightarrow y$ for which $\exists z \in x : support\ x \rightarrow y = support\ (x \rightarrow z \rightarrow y)$ [30]. Furthermore, Bayardo et al. [8] define a more dominant minimum improvement constraint in order to discard the redundant rules. The improvement of rule $x \rightarrow y$ is defined as improvement $(x \rightarrow y) = confidence\ (x \rightarrow y) - \max_{z \subset x}(confidence(x \rightarrow y))$. In the datasets where there is a predefined class label (i.e. classification tasks), frequent pattern mining can contribute to discovering strong associations between occurring attribute and class values [22]. In [11] the potential usage of frequent pattern mining for classification problems was investigated and successfully applied to the problem. Their approach discovered classification rules by directly discovering the frequent patterns from the datasets with predefined class labels. The results reported were promising since the discovered knowledge model had high accuracy and efficiency for the classification problem.

In the work presented in this paper, we focus on ascertaining the frequent patterns that have been extracted from a relational database, and that satisfy the minimum transaction-based support and confidence thresholds.

Let us denote the set of these frequent patterns as *FP*. One of the attributes from the data is considered as a class to be predicted for classification task purposes. Hence, we consider only consider those patterns from *FP* that contain this class attribute, as they will represent the set of values that frequently occur together when a particular class value is present. Hence, as such these patterns can be seen to have predictive power and can be evaluated for their accuracy on correctly predicting the class value from the trained data (classification accuracy), and unseen data (predictive accuracy).

In addition to predictive accuracy, simple rules are preferred as they are easier to comprehend and are expected to perform better on unseen data since they are more general. Hence, during the process of optimizing a rule set, a trade-off needs to be made between several factors, the common ones are:

- *Misclassification rate* (*MR*)-number of incorrectly classified instances
- *Coverage rate*(*CR*)-number of captured instances
- *Generalization power* (*GP*)-capability of correctly classifying future instances.

When optimizing the rule set, the *MR* should be minimized while the *CR* should be maximized. Good *GP* is achieved by simplifying the rules in terms of overall rule set size and the number of attribute constraints in the rule. The trade-off occurs especially when the dataset is characterized by continuous attributes where a valid attribute range constraint needs to be determined for a particular rule. Increasing the range constraint usually leads to the increase in *CR* of that rule but at the cost of an increase in *MR* of that rule. Similarly, if the rules are too general, they may lack the specificity to distinguish some domain characteristics and hence the *MR* would increase. Generally speaking, an optimized rule set should be either more accurate than the original rule set and/or the balance between the trade-off factors should be much greater. For example, if there are many rules with small *CR* but very low *MR*, a rule set with a significantly smaller number of rules may be preferred even at the cost of an increase in *MR*.

Since the number of patterns/association rules generated through association rule mining can be quite large, their usefulness for a classification/prediction task may be limited unless they are significantly reduced in size and number. While their *MR* may be small, their *GP* is likely to be poor as all frequent patterns are considered, and these can be insignificant, redundant and unnecessarily complex. Hence in this paper, we aim to apply a variety of statistical/heuristic methods to reduce the pattern/rule set size and simplify individual rules.

Let us denote the patterns from the frequent set *FP* that have a class label (value), as *FPC*. The problem can be stated as: given *FPC* with accuracy *ac*, reduce *FPC* into *FPC*' such that *FPC*' has accuracy $\geqslant (ac - \varepsilon)$, where $\varepsilon$ is an arbitrary user defined small value ($\varepsilon$ is used to reflect the noise that is often present in real world data).

## 3. Related work

Measuring interestingness of knowledge patterns is an active and important area of data mining research. Different methods have been proposed for discovering interesting rules from data and these can be generally categorized into three main classes, namely objective, subjective and semantic measures [6,14,17,24]. However, the measures usually reflect just the usefulness of rules with respect to the specific database being observed [30]. The data mining approaches consider the whole search space to find all possible pattern/rules satisfying specific criteria (i.e. association rules). While these criteria, impose some constraints on the discovery of strong patterns/rules, many spurious, misleading, uninteresting and insignificant rules in those domains may still be produced [17]. This problem arises because some association rules are discovered due to pure coincidence resulting from a certain randomness in the particular dataset being analyzed. This is further supported by Lallich et al. [20] who asserts that the patterns discovered using the traditional association rule mining framework can be either a true discovery or merely an artefact of random selection.

Statistics have previously addressed the issue of how to separate out the random effects to determine whether the measured association (or difference in other areas) is significant [4,18].

However, the statistical significance assessment is still poorly understood and remains one of the challenging data mining problems to solve [19].

To date, works proposed by Hämäläinen and Nykänen [16], Kirsch et al. [19], Lallich et al. [20], Webb [29,30] and Weiß [32] recognized the need for a statistically significant pattern. Webb [29,30] demonstrate the capabilities of their approach by performing two techniques namely, the holdout and direct adjustment, to check for a productive and significant rule. This approach was motivated to extend the works done by Bay and Pazzani [7], Meggido and Srikant [25] to strictly control the false discovery, which is the error of rejecting a null hypothesis and thus falsely accepting a pattern. The initial work on avoiding false discovery has been successfully addressed by Meggido and Srikant [25], but the method applied is valid only when applied to sparse data transactions. [19] have successfully developed a novel methodology to identify a meaningful support threshold for a given dataset, in order to control the false discovery rate. This technique differentiates the significant itemsets with a small false discovery rate as those itemsets that deviate substantially from the expected random dataset. Moreover, Lallich et al. [20] proposed and utilized a bootstrap-based method to control the multiple risks and avoid the risk of false discovery. While another work proposed by Wei et al. [31] namely the *support-match* framework, was considered capable of mining effective rules and also pruning the low relation rules. Furthermore, AydIn and Güvenir [6] proposed an *association rule set stream* in generating and measuring the interestingness of steam data that vary in size and change over time.

Weiß [32] proposed a measure to express the trustworthiness of the association rule based on precision values which rely on the estimator of confidence interval. Both precision measures provided sufficient, reliable and highly predictive rules. Another notable work was proposed in [16], who implemented the *StatApriori* algorithm which searches statistically significant and non-redundant rules. This algorithm was developed to control the existence of false negatives and false positives discovered in association rule mining.

Thus, additional measures based on statistical independence and correlation analysis are often required to ensure that the results have a sound statistical basis and are not purely the result of random coincidence. The statistical approach offers a reliable way of identifying significant rules that are statistically valid.

## 4. Proposed method

The motivation behind our proposed method is to investigate how data mining and statistical measurement techniques can be combined to arrive at a more reliable and interesting set of rules. Generally speaking, we interpret interesting rules as those rules that have a sound statistical basis and are not redundant. Such an approach requires a sampling process, hypothesis development, model building and finally a measurement using statistical analysis techniques to verify and ascertain the usefulness and quality of the rules discovered. This will filter out the redundant, misleading, random and coincidentally occurring rules, while at the same time ensuring the accuracy of the rule set.

### 4.1. Conceptual framework

Fig. 1 shows the proposed framework. The formal definition of the conceptual framework is defined as follows:

**Definition 1** (*Relational database*). Given a relational database $D$, $I = \{i_1, i_2, \ldots, i_k\}$ the set of distinct items in $D$, and $C = \{c_1, c_2, \ldots, c_k\}$ the set of class labels in $D$. Assume that $D$ contains a set of $n$ instances $D = \{x_i, y_i\}_{i=1}^{n}$, where $x_i \subseteq I$ is a set of items and $y_i \in C$ is a class label. The training dataset $D_{tr} \subseteq D$ and the testing dataset $D_{ts} \subseteq D$.

**STEP 1**. The preprocessing is applied to each $x_i$ in $D_{tr}$ in order to obtain clean and consistent data. These preprocessing techniques include the removal of missing values and discretization of attributes with continuous values.

**STEP 2**. We determine the relevance of input attributes $x_i$ by ascertaining their importance in predicting the class label $y_i$ in $D_{tr}$ using a statistical-heuristic measure. Any irrelevant attributes are removed from the dataset, and we represent the filtered database as $\widetilde{D_{tr}}, \widetilde{I} \subseteq I$.

A powerful technique for this purpose is the Symmetrical $\tau$ [35] which is a statistical-heuristic feature selection criterion. It measures the capability of an attribute to predict the class of another attribute. Let there be $R$ rows and $C$ columns in the contingency table for two attributes $x_i$ and $y$. The probability that an individual belongs to row category $r$ and column category $c$ is represented as $P(rc)$, and $P(r+)$ and $P(+c)$ are the marginal probabilities in row category $r$ and column category $c$, respectively. The measure is based on the probabilities of one attribute value occurring together with the value of the second attribute. In this sense, the $y$ attribute can be seen as a representative of the class attribute, and the Symmetrical $\tau$ measure for the capability of input attribute in predicting the class attribute is defined as [35].

$$\tau(x_i, y) = \frac{\sum_{c=1}^{C}\sum_{r=1}^{R}\frac{P(rc)^2}{P(+c)} + \sum_{r=1}^{R}\sum_{c=1}^{C}\frac{P(rc)^2}{P(r+)} - \sum_{r=1}^{R}P(r+)^2 - \sum_{c=1}^{C}P(+c)^2}{2 - \sum_{r=1}^{R}P(r+)^2 - \sum_{c=1}^{C}P(+c)^2} \quad (1)$$

The higher values of the Symmetrical $\tau$ measure would indicate better discriminating criteria (features) for the class that is to be predicted in the domain. Symmetrical $\tau$ has many more desirable properties in comparison with other feature selection techniques, as reported in [35].

In Section 5.1, we evaluate the capabilities of Symmetrical $\tau$ as the determinant of the relevance of attributes by comparing it with an information-theoretic measure. The information-theoretic measures are principally comprehensible and useful since they can be interpreted in terms of information. For a rule interestingness measure, the relation is interesting when the antecedent provides a great deal of information about the consequent [10]. Although several information-theoretic measures exist, we compared only Symmetrical $\tau$ with Mutual Information measurement technique which is the most well known of the techniques. The Mutual Information is based on information theory to evaluate rules. This approach describes how much information one random variable imparts about another one [23]. The definition of Mutual Information is based on [28].

$$M(x_i, y) = \frac{\sum_r \sum_c P(x_i, y) \log \frac{P(x_i, y)}{P(x_i)P(y)}}{\min(-\sum_r P(x_i) \log P(x_i) - \sum_c P(y) \log P(y))} \quad (2)$$

The information that $y$ tells us about $x_i$ is the reduction in uncertainty about $x_i$ due to knowledge of $y$. The greater the values of $M$, the more information $x_i$ and $y$ contain about each other [23]. The Symmetrical Tau features selection technique is utilized in our approach to provide the relative usefulness of attributes in predicting the value of the class attribute, and discard any of the attributes whose relevance value is fairly low. This would prevent the generation of rules which then would need to be discarded anyway once it were found that they include irrelevant attributes.

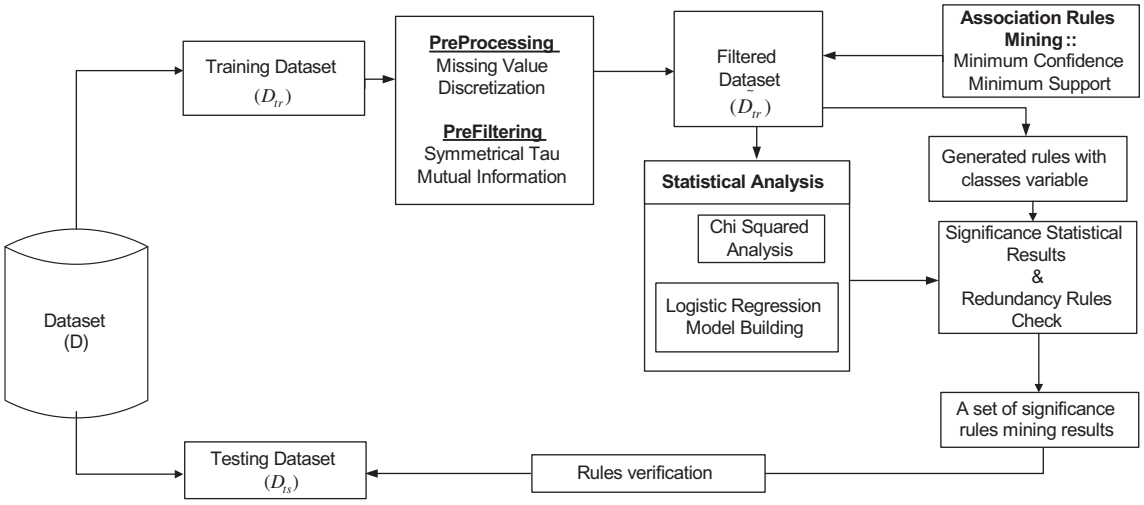


**Fig. 1.** Proposed framework for rule interestingness analysis.

**Table 1**
Comparison between ST and MI for Adult dataset (initial data proportion).

| # of Values | Variables | ST values | # of Values | Variables | MI values |
|---|---|---|---|---|---|
| 7 | Marital status | 0.1448 | 6 | Relationships | 0.1662 |
| 6 | Relationship | 0.1206 | 7 | Marital status | 0.1575 |
| 6 | Capital gain | 0.0706 | 16 | Education | 0.0934 |
| 8 | Education number | 0.0688 | 14 | Occupation | 0.0932 |
| 16 | Education | 0.0528 | 8 | Education number | 0.0900 |
| 2 | Sex | 0.0470 | 10 | Age | 0.0894 |
| 14 | Occupation | 0.0469 | 10 | Hours per week | 0.0545 |
| 10 | Age | 0.0432 | 6 | Capital gain | 0.0475 |
| 5 | Capital loss | 0.0361 | 2 | Sex | 0.0374 |
| 10 | Hours per week | 0.0354 | 5 | Capital loss | 0.0238 |
| 7 | Work class | 0.0166 | 7 | Work class | 0.0171 |
| 5 | Race | 0.0085 | 41 | Native country | 0.0093 |
| 41 | Native country | 0.0077 | 5 | Race | 0.0083 |
| 10 | FNLWGT | 0.0002 | 10 | FNLWGT | 0.0002 |

**STEP 3**. For a given $\widetilde{D_{tr}}$, we discover all association rules based on *minimumsupport*and *minimumconfidence* Apriori framework. Since we are dealing with a classification problem, we consider only those rules that contain a class label, and hence we formulize rules as follows:

**Definition 2** (*Frequent rule*). A rule is defined by $x_i \rightarrow y_i$ where $x_i$ is the antecedent and $y_i$ the consequent. An initial frequent rule $F$ is a subset of a record from the filtered database $\widetilde{D_{tr}}$ that satisfy both *minimumsupport* and *minimumconfidence* threshold.

**STEP 4a**. For a given $\widetilde{D_{tr}}$, the occurrence of $x_i$ is independent of the occurrence of $y$ if $P(x_i \cup y) = P(x_i)P(y)$; otherwise $x_i$ and $y$ are dependent and correlated [17]. We measure the correlation between $x_i$ and $y$ as follows:

$$corr(x_i, y) = \frac{P(x_i \cup y)}{P(x_i)P(y)} \tag{3}$$

For a given correlation value based on Eq. (3), we use the $\chi^2$ statistic value to determine if the correlation is statistically significant [17].

**STEP 4b**. For a given $\widetilde{D_{tr}}$, we develop several logistic regression models. We select the model that fits the data well and the model with the highest predictive capabilities. We denote the selected model as $\ln(\widetilde{y})$.

**Definition 3** (*Logistic regression model*).

$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i + \varepsilon$,
where;
$\ln(y)$ = Natural logarithm of the odds ratio,
$\beta_0 + \beta_1 + \cdots \beta_i$ = Coefficients of the input variables,
$\varepsilon$ = Error variable,
$y$ = Dichotomous class attribute,
$x_i$ = Input attributes.

We calculate the log likelihood value based on Eq. (4) to estimate the attribute's coefficient. We then use the statistical hypothesis to determine whether the input attributes are significantly related to class attribute.

$$L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{4}$$

**STEP 5**. Let $F$ be a set of initial frequent rules generated from Step 3. We check the significance of $F$ generated from $\widetilde{D_{tr}}$ by verifying $F$ with statistical analysis in Steps 4a, 4b and redundancy check (refer to Section 2).

**Definition 4** (*Significant rules*). Let $F$ denote the set of rules with an associated target, $F_i(x_i \rightarrow y_i) \in \widetilde{D_{tr}}$

(4a) Chi squared test discard rules $F_i(x_i \rightarrow y_i)$ for which $\exists x_i \in \widetilde{D_{tr}} : \chi^2$ value is not significant.

(4b) Logistic regression $\widetilde{\ln(y)}$ discards rules $F_i(x_i \rightarrow y_i)$ for which $\exists x_i \in \widetilde{D_{tr}} : \beta_i x_i$, value is not significant. We denote the Frequent rules based on Statistical Analysis in 4a and 4b as FSA.

(4c) Minimum improvement redundant rule constraints discard rule $FSA_i(x_i \rightarrow y_i) = confidence\ x_i \rightarrow y_i - \max_{z_i \subset x_i}\ confidence\ (z_i \rightarrow y_i))$ [8].
We denote the frequent rules based on minimum improvement redundant rule constraints as $\widetilde{F}$.

**STEP 6**. For each $F$, $FSA$ and $\widetilde{F}$, we calculate the rule's accuracy by verifying with $D_{tr}$ and $D_{ts}$. The process as a whole is summarized in Tables 2 and 3.
The accuracy of rules is calculated with respect to the correctly classified instances from both $D_{tr}$ (i.e. classification accuracy) and $D_{ts}$ (i.e. predictive accuracy).

The combination of these rule ascertaining strategies will facilitate the association rule mining framework to determine the right and high quality rules. These rules will have a sound statistical basis and we can be more confident that they reflect the real world situation.

**Algorithm 1**: Significant association rule mining

**Input**: database $D$ with class attribute, *minimum support* and *minimum confidence* values

**Output**: A set of significant rule $(\widetilde{F})$

1. Divide the database $D$ into $D_{tr}$ and $D_{ts}$,
2. Calculate and rank the importance of input attributes $x_i$ toward the class attribute in $D_{tr}$ based on Symmetrical $\tau$. Discard $x_i$ which are less relevant and denote the filtered database as $\widetilde{D_{tr}}$.
3. Identify the non-significant $x_i$ in $\widetilde{D_{tr}}$ based on statistical analysis.
   - 3.1 Calculate the $\chi^2$ for each $x_i$ in $\widetilde{D_{tr}}$ and identify $x_i$ which are not significantly correlatesd with class attribute,
   - 3.2 Develop and fit several Logistic regression models from $\widetilde{D_{tr}}$. Choose a model that fits the data and has the highest predictive capabilities, denote the selected model as $(\widetilde{\ln(y)})$. Estimate the coefficient of $x_i$ in $\widetilde{\ln(y)}$ and identify $x_i$ which are not significantly correlated to class attribute.
4. Generate a frequent rule set $F$ from $\widetilde{D_{tr}}$ based on *minimumsupport* and *minimumconfidence* Apriori framework.

For each $x_i$ in $F \in \widetilde{D_{tr}}$,
for each $\chi^2$ value in $x_i \in \widetilde{D_{tr}}$,
if $\chi^2$ is not significant, then discard any $F$ that contains $x_i$ else for each $\beta_i x_i$ value in $\widetilde{\ln(y)}$
where $x_i \in \widetilde{D_{tr}}$
if $\beta_i x_i$ not significant then discard any $F$ that contains $x_i$ denote this as $FSA$,
if any $FSA$, failed the *Minimum Improvement Redundant Check*
(Definition 4c) then discard $FSA$, else retain $FSA$.
Return a set of retained $FSA$, denoted as $\widetilde{F}$.

## 5. Experimental results

The evaluation of the unification framework is performed using the Adult, Iris and Wine dataset, which are real world datasets of varying complexity obtained from the UCI Machine Learning Repository. We employed an efficient breadth-first Apriori based algorithm [3] for generating candidate association rules. Since all the datasets used are supervised which reflects a classification problem, we have chosen the target variable as the right hand side/consequence of the association rules discovered during association rule mining analysis. In Section 5.1, we first compare two established measures for feature selection, namely Symmetrical $\tau$ and Mutual Information to measure the capability of attributes in predicting the values of the target attribute. We then discuss the effect that the occurrence of rare target data (Section 5.2) and unified target data (Section 5.3) in a dataset can have on the proposed framework. Finally, in Section 5.4 we discuss the performance of the framework as a whole when evaluated on all three datasets.

As for many real world problems, the forms of the input and target attributes emerge from a wide range of measurement levels. In handling these types of attributes, we apply the binning approach to improve the boundary of the continuous variables. These bounds are created to reflect the upper and lower values for the input variables [12]. For all continuous attributes in Adult, Iris and Wine, we apply equal depth binning approach methods. This approach ensures that we have a manageable data size by reducing the number of distinct values per attributes [17]. Other discrete attributes in the Adult dataset were preserved in their original state. In order to gauge the effect of rules accuracy on different sets of partitioning for each dataset, $k$-fold cross validation approach was utilized, to ensure that we obtained relatively low bias and variance [17].

### 5.1. Comparing Symmetrical $\tau$ (ST) with Mutual Information (MI)

ST and MI are capable of defining irrelevant attributes, but they are different from each other in terms of their approach as aforementioned in Section 4.1. We apply both ST and MI to the Adults, Wine and Iris dataset.

Throughout the experiments, we found that the MI approach favors variables with more values. This observation is in accord with [10]. Conversely, the procedure based on ST produces a more stable selection of variables which does not favor any specific variables criterion. This is in agreement with the claim in [35], that ST is fair in handling multi-valued variables.

The comparison results for the Adult dataset is shown in Table 1, where the capabilities of attributes in predicting the values of attribute 'Income' ( <=50 K and >50 K) are measured. In Section 5.4, we discuss in detail the selection of relevant attributes within the framework as a whole.

### 5.2. Rare target data problems

Table 2 shows 3 experiments performed for the Adult dataset. For the Adult dataset, we have limited the consequent of the rules to be either Income <=50 K or Income >50 K, as these reflect the possible values for the chosen target attribute (i.e. Income). The initial proportion of these target values is unbalanced, producing an infrequent target value in the data for the Adult dataset (i.e. >50 K) (rare target data).

Initially, we apply the Apriori algorithm to discover association rules based on the initial data proportion of the training dataset, and the results are shown in the second row of Table 2. We then attempted to balance the dataset so that the number of records occurring with target value ' >50' is equal to the number of records occurring with the target value ' <=50 K'. The third row shows the results when we have reduced the training dataset by removing some of the records with the more frequent target value (i.e. <=50 K).

Finally, we replicated several records in the training dataset (row 4 of Table 2). This replication process generated additional

**Table 2**
Rules accuracy for Adult data.

| Experimental approaches | Dataset description | Rule # | Type of analysis | Accuracy | |
|---|---|---|---|---|---|
| | | | | Classification (%) | Prediction (%) |
| Initial proportion | Training: 30,162 records $\begin{bmatrix} <= 50K : 22,654 \\ > 50K : 7508 \end{bmatrix}$ | 164 | Initial rules | 86.75 | 86.87 |
| | | 53 | Statistical analysis | 87.73 | 87.92 |
| | | 42 | Redundancy check | 87.99 | 88.13 |
| | Test: 15,060 records | | | | |
| Balance data | Training: 15,016 records $\begin{bmatrix} <= 50K : 22,654 \approx 7508 \\ > 50K : 7508 \Rightarrow 7508 \end{bmatrix}$ | 421 | Initial rules | 71.55 | 60.56 |
| | | 51 | Statistical analysis | 73.87 | 58.28 |
| | | 30 | Redundancy check | 74.00 | 63.80 |
| | Test: 15,060 records | | | | |
| Replication data | Training: 45,178 records $\begin{bmatrix} <= 50K : 22,654 \Rightarrow 22,654 \\ > 50K : 7508 \approx 7508 \times 3 = 22,524 \end{bmatrix}$ | 255 | Initial rules | 71.65 | 59.86 |
| | | 51 | Statistical analysis | 73.64 | 58.28 |
| | | 32 | Redundancy check | 73.61 | 61.70 |
| | Test: 15,060 records | | | | |

**Table 3**
Rules accuracy for Wine and Iris data.

| Dataset name | Dataset description | Rule# | Type of analysis | Accuracy | |
|---|---|---|---|---|---|
| | | | | Classification (%) | Prediction (%) |
| Wine | Train: 107 records | 195 | Initial rules | 87.53 | 79.44 |
| | Test: 71 records | 17 | Statistical analysis | 85.07 | 81.98 |
| | | 16 | Redundancy check | 85.07 | 81.98 |
| Iris | Train: 90 records | 52 | Initial rules | 92.86 | 90.99 |
| | Test: 60 records | 22 | Redundancy check | 88.15 | 85.29 |

records for training data so that any value from the set of target values has a more similar frequency of occurrence in the training dataset and this will represent a similar proportion between each target item/value. By applying the Apriori algorithm on both balanced and replicated training data, a large volume of rules has been generated in comparison with the initial data proportion of Adult dataset. Yet, with a proper statistical analysis and redundancy check within all designed experiments, we managed to reduce the initial rule set by at least 75%.

### 5.3. Unified target data

Table 3 shows the results of applying the proposed framework on the Iris and Wine dataset, which represent unified target data. For both datasets, the generation of rules was based on an Apriori algorithm. The initial rules for Wine dataset were 195. Based on the statistical analysis, we managed to reduce the rule set to contain only 17 rules, and finally, with the application of redundancy check, we found that only 8% (16 rules) of 195 rules were significant.

For the Iris dataset, the initial rule set was 52 and we reduced it to 22 significant rules based on the redundancy check (as the statistical analysis did not consider any attributes as irrelevant). Through the statistical analysis and redundancy check, we managed to discard a high number of spurious and insignificant association rules generated from both Wine and Iris datasets.

### 5.4. Overall framework performance

Using the whole dataset as input would produce a large number of rules, many of which are created by the presence of irrelevant attributes. Since the ST has more advantageous properties in comparison to MI, the ST feature selection criterion was used earlier in the process to remove any irrelevant attributes. This would prevent the generation of rules that include any irrelevant attributes. Hence, in this experiment it is not necessary to use ST to further verify the rules as the rules were created from the attribute subset considered as relevant according to the measure, as was done in [26,27]. The attributes were ranked according to their decreasing ST and a relevance cut-off point was chosen. In this experiment, the cut off value was selected based on the significant difference between the ST values in decreasing order. The significant difference was considered to occur in the ranking at the position where that attribute's ST value is less than half of the previous attribute's ST value in the ranking. At this point and below in the ranking, all attributes are considered as irrelevant. For example, for the Adult dataset results presented in Table 1, the relevance cutoff value is 0.0166. This is due to the ST value of attribute 'Hours per week' being more than double of the ST value for attribute 'Work class'. Thus, the subset of data now consists of 10 attributes: Marital status, Relationship, Capital gain, Education number, Education, Sex, Occupation, Age, Capital loss and Hours per week.

We proceed with the application of the association rule mining algorithm and verification of the extracted rules through statistical analysis. On the right hand side of Tables 2 and 3, we show the progressive difference in the number of rules generated as statistical analysis and redundancy checks are being utilized. We also show the respective classification (% of correctly classified instances from the training set) and predictive accuracy (% of correctly classified instances from the training set) of those rule sets. For most of the discovered rules, the classification accuracy was consistently higher than for the predictive accuracy. This is due to the fact that the discovered rules were generated from the training set, and as a consequence, the rules would have fitted well with all the criteria that exist predominantly in the training set.

By analyzing the results of the application of the proposed method (rightmost column of Tables 2 and 3), we can see that the combination of statistical significance analysis and redundancy analysis provided an effective means of discarding non-significant rules, Furthermore, this was not always at a the cost of a significant reduction in the accuracy as is discussed next.

For the Adult dataset, more than 75% of the rules have been discarded. On average, the rule's accuracy either is in balance or the replication based approach is lower compared to their initial data proportion. Thus, for this dataset, balancing and replicating the

records will decrease the classification and predictive accuracy. When evaluating the difference in accuracy of the rule set due to the application of the proposed method, the following can be observed. When the rule set was reduced based upon the statistical analysis and the redundancy check, the predictive accuracy actually increased in comparison to the predictive accuracy of the whole rule set.

As for the Wine and Iris dataset, at least 60% of rules have been discarded from the original discovered rules set. However, to some extent, the rule's accuracy of each of these two datasets has some interesting differences. We noted that none of the input attributes in the Iris dataset was discarded based on the statistical analysis approach either by the Chi squared or Logistic regression. Each of the four input attributes in the Iris dataset were statistically significant in predicting the target attributes. Due to this, there is some deterioration in rules accuracy for the Iris data compared to the Wine dataset. For the Iris dataset, 30 rules were removed but the accuracy was reduced by about 4–5%. On the other hand, for the Wine dataset, 179 rules were removed and while there was a slight decrease in classification accuracy, the predictive accuracy has actually increased. This demonstrates the importance of ascertaining the association rules by statistical means, as in contrast to the Iris dataset, the simplified rule set for the Wine dataset is more general and performs better on unseen data. The reduction in classification accuracy could have been caused, by the fact that many of the extracted rules reflect the associations found in the sample data and not with respect to the whole data distribution.

Overall, the results highlight the importance of ascertaining the association rules by both statistical analysis and redundancy check, as for both Adult and Wine datasets, the simplified rule set is more general and has higher predictive accuracy than the more specific initial rule set.

## 6. Conclusions and future works

This paper has presented a framework which integrates a number of ways for ascertaining the extracted data mining rules. The aim was to ascertain the quality of the rules discovered from association rule mining which has a huge amount of rules and complex attributes measurement levels with an integrated statistical and heuristic measurement technique. The experimental results show that this framework managed to reduce a large number of non-significant and redundant rules while at the same time preserving relatively high accuracy. This indicates the potential of the framework to provide significant rules when applied to the structured or relational data. As part of our ongoing work, we intend to use the proposed framework to ascertain more complex rules which are discovered from semi-structured data.

## References

[1] C.C. Aggarwal, P.S. Yu, A new framework for itemset generation, in: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Seattle, Washington, United States, 1998.
[2] R. Agrawal, T. Imieliski, A. Swami, Mining association rules between sets of items in large databases, SIGMOD Rec. 22 (1993) 207–216.
[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Databases, 1994.
[4] A. Agresti, An Introduction to Categorical Data Analysis, second ed., Wiley-Interscience, Hoboken, N.J, 2007.
[5] Y. Aumann, Y. Lindell, A statistical theory for quantitative association rules, J. Intell. Inf. Syst. 20 (2003) 255–283.
[6] T. AydIn, H.A. Güvenir, Modeling interestingness of streaming association rules as a benefit-maximizing classification problem, Knowledge Based Syst. 22 (2009) 85–99.
[7] S.D. Bay, M.J. Pazzani, Detecting group differences: mining contrast sets, Data Min. Knowledge Discovery 5 (2001) 213–246.
[8] R.J. Bayardo, R. Agrawal, D. Gunopulos, Constraint-based rule mining in large, dense databases, Data Min. Knowledge Discovery 4 (2000) 217–240.
[9] L. Bing, H. Wynne, M. Yiming, Mining Association Rules with Multiple Minimum Supports, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, United States, 1999.
[10] J. Blanchard, F. Guillet, R. Gras, H. Briand, Using information-theoretic measures to assess association rule interestingness, in: Proceedings of the 5th IEEE International Conference on Data Mining, 2005.
[11] H. Cheng, X. Yan, J. Han, P.S. Yu, Direct Discriminative Pattern Mining for Effective Classification, Piscataway, NJ, USA, 2008. pp. 169–178.
[12] T.S. Dillon, T. Hossain, W. Bloomer, M. Witten, Improvements in supervised brainne: a method for symbolic data mining using neural networks, in: S. Spaccapietra, F.J. Maryanski (Eds.), IFIP TC2/WG2.6 Seventh Conference on Database Semantics (DS-7), Leysin, Switzerland, 1998, pp. 67–88.
[13] A.A. Freitas, On rule interestingness measures, Knowledge Based Syst. 12 (1999) 309–315.
[14] L. Geng, H.J. Hamilton, Interestingness measures for data mining: a survey, ACM Comput. Surv. 38 (2006) 9.
[15] A. Goodman, C. Kamath, V. Kumar, Data analysis in the 21st century, Stat. Anal. Data Min. 1 (2008) 1–3.
[16] W. Hämäläinen, M. Nykänen, Efficient discovery of statistically significant association rules, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, 2008.
[17] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.
[18] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, Wiley, New York, 1989.
[19] A. Kirsch, M. Mitzenmacher, A. Pietracaprina, G. Pucci, E. Upfal, F. Vandin, An efficient rigorous approach for identifying statistically significant frequent itemsets, in: Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Providence, Rhode Island, USA, 2009.
[20] S. Lallich, O. Teytaud, E. Prudhomme, Association rule interestingness: measure and statistical validation, Qual. Meas. Data Min. (2007) 251–275.
[21] N. Lavrač, P. Flach, B. Zupan, Rule Evaluation Measures: A Unifying View, Inductive Logic Programming. pp. 174–185.
[22] J. Li, H. Shen, R.W. Topor, Mining the optimal class association rule set, Knowledge Based Syst. 15 (2002) 399–405.
[23] S. Lotfi, M.H. Sadreddini, Mining Fuzzy Association Rules Using Mutual Information, International MultiConference of Engineers and Computer Scientists Publishing, Hong Kong, 2009.
[24] K. McGarry, A survey of interestingness measures for knowledge discovery, Knowl. Eng. Rev. 20 (2005) 39–61.
[25] N. Meggido, R. Srikant, Discovering predictive association rules, in: 4th International Conference on Knowledge Discovery in Databases and Data Mining, Publishing, 1998, pp. 274–278.
[26] I.N. Mohd. Shaharanee, F. Hadzic, T. Dillon, Interestingness of association rules using symmetrical tau and logistic regression, in: A. Nicholson, X. Li (Eds.), AI 2009, 2009, pp. 442–431.
[27] I.N.M. Shaharanee, T.S. Dillon, F. Hadzic, Ascertaining association rules using statistical analysis, in: P.S. Sandhu (Ed.), 2009 International Symposium on Computing, Communication and Control (ISCCC 2009), Singapore, 2009, pp. 180–188.
[28] P.-N. Tan, V. Kumar, J. Srivastava, Selecting the right interestingness measure for association patterns, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002.
[29] G.I. Webb, Preliminary investigations into statistically valid exploratory rule discovery, in: Australasian Data Mining Workshop (AudDM03), 2003, pp. 1–9.
[30] G.I. Webb, Discovering Significant Patterns, Machine Learning, Springer, 2007. pp. 1–33.
[31] J.-M. Wei, W.-G. Yi, M.-Y. Wang, Novel measurement for mining effective association rules, Knowledge Based Syst. 19 (2006) 739–743.
[32] C. Weiß, Statistical mining of interesting association rules, Stat. Comput. 18 (2008) 185–194.
[33] H. Yun, D. Ha, B. Hwang, K.H. Ryu, Mining association rules on significant rare data using relative support, J. Syst. Softw. 67 (2003) 181–191.
[34] H. Zhang, B. Padmanabhan, A. Tuzhilin, On the discovery of significant statistical quantitative rules, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA, 2004.
[35] X.J. Zhou, T.S. Dillon, A statistical-heuristic feature selection criterion for decision tree induction, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 834–841.